Applied Data Science
Project Update Presentation:

Best City for investing a Chinese Restaurant

...

Nicolas Fan
04/25/2019

Project Inspiration

Suppose an investor comes to consult me. He has a budget of **250,000 dollars** and wants to invest in a **Chinese restaurant** in North America. Which city would be the best choice and why?

Yelp Open Dataset

The dataset contains:

6,685,900 reviews

192,609 businesses

200,000 pictures

10 metropolitan areas

The files:

- **business.json**
- **review.json**
- user.json

What's inside business.csv?

1. Business_id
2. Name
3. Neighborhood
4. Address
5. City
6. State
7. Postal_code
8. Latitude
9. Longitude
10. Stars
11. Review_count
12. Is_open

1. c7X2SdKxVJMaOnFROO8WEg
2. "Finga Lickin' Caribbean Eatery"
3.
4. "2838 The Plz"
5. Charlotte
6. NC
7. 28205
8. 35.236823
9. -80.801084
10. 4.5
11. 1
12. Pizza;Food;Caribbean

Logic & Assumption

The following indicators can indirectly or directly indicate whether Chinese food are popular in a region.

- **Open rate**
- **Average Star**
- **Review counts**

Data Cleaning

Step 1: Chinese diner

```
diner_chinese =
(df_GTA.loc[df['categories
'].isin(['Chinese'])])
```

Step 2: define region

```
GTA =
(df.loc[df['city'].isin(['Ajax',
'Brampton', 'Burlington',
'Markham','Toronto','Missi
ssauga', 'Newmarket',
'Oakville', 'Oshawa',
'Pickering', 'Richmond
Hill', 'Vaughan',
'Whitby'])])
```

Step 3: irrelevant info

```
columns_needed =
['business_id', 'name',
'neighborhood', 'address',
'city', 'state', 'latitude',
'longitude', 'stars',
'review_count', 'is_open']
```

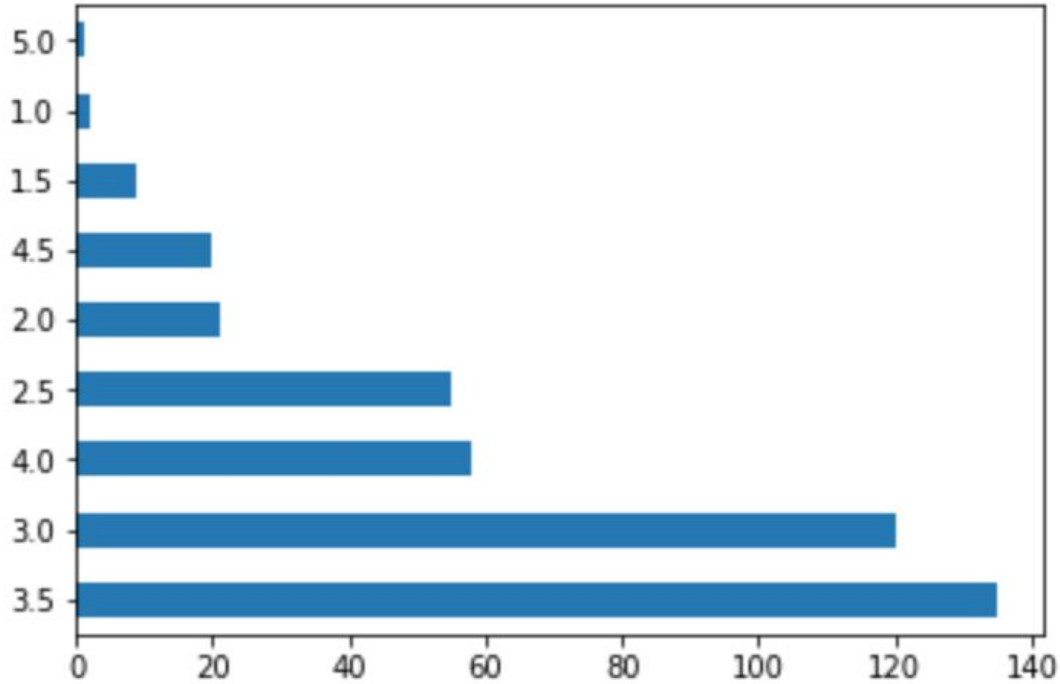
Analysis

	state	postal_code	latitude	longitude	stars	review_count	is_open
248	ON	M5T 1G6	43.653324	-79.395372	3.5	165	1
672	ON	L5M 5S5	43.558921	-79.742916	3.0	14	1
1329	ON	L4Z 3K8	43.614583	-79.662815	3.5	41	1
1467	ON	M6R 1A7	43.639365	-79.442530	3.0	8	0
2073	ON	M1S 2B7	43.786989	-79.276122	3.5	24	1
3534	ON	L3P 5J5	43.876709	-79.285820	3.5	118	1
3815	ON	L3T	43.819342	-79.399660	3.0	9	1
3924	ON	L4B 3N7	43.866086	-79.387103	2.5	3	0
4518	ON	L3Y 8S3	44.046096	-79.437159	3.0	7	1
5091	ON	M2H 2N5	43.805515	-79.337624	3.0	3	1
5161	ON	L6L 5B3	43.433704	-79.702300	3.5	4	0
5456	ON	L3R 9Y7	43.846579	-79.357335	3.0	6	1
5732	ON	L6H 6W5	43.476373	-79.726544	3.0	11	1

	address	city
248	"421 Dundas St W, 3rd Fl"	Toronto
672	"5602 Tenth Line W, Unit 110"	Mississauga
1329	"Sandalwood Mall, 30 Bristol Road E, Unit 3"	Mississauga
1467	"1533 Queen Street W"	Toronto
2073	"3 Glen Watford Drive"	Toronto
3534	"1 Raymerville Dr"	Markham
3815	"300 John Street"	Markham
3924	"9425 Leslie Street"	Richmond Hill
4518	"869 Mulock Drive, Unit 12"	Newmarket
5091	"3560 Victoria Park Avenue"	Toronto
5161	"649 Fourth Line"	Oakville
5456	"8368 Woodbine Ave"	Markham
5732	"2530 Sixth Line, Unit 13"	Oakville

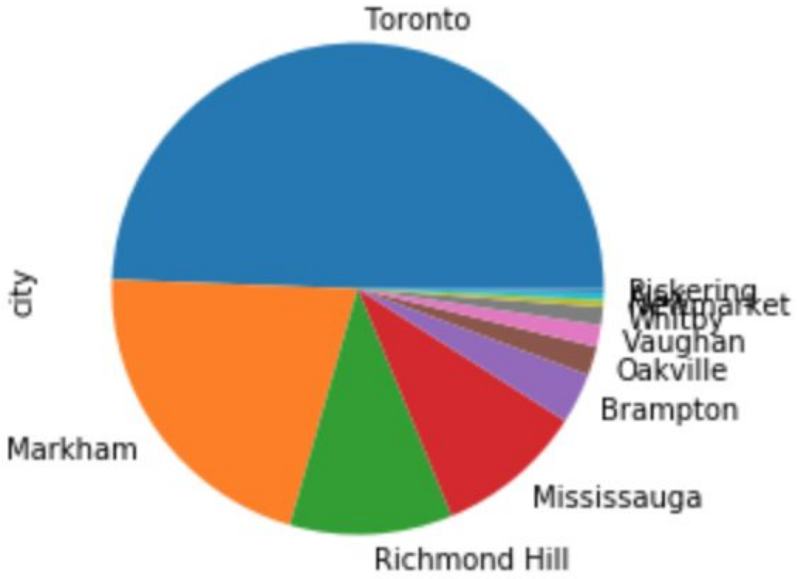
[16]:
```
#Among the many Chinese restaurants in Toronto, we clearly observed that
#As can be seen from the figure, the Chinese restaurant's score on the y
GTA_chinese['stars'].value_counts().plot(kind='barh')
```

t[16]: <matplotlib.axes._subplots.AxesSubplot at 0x127a4eeb8>



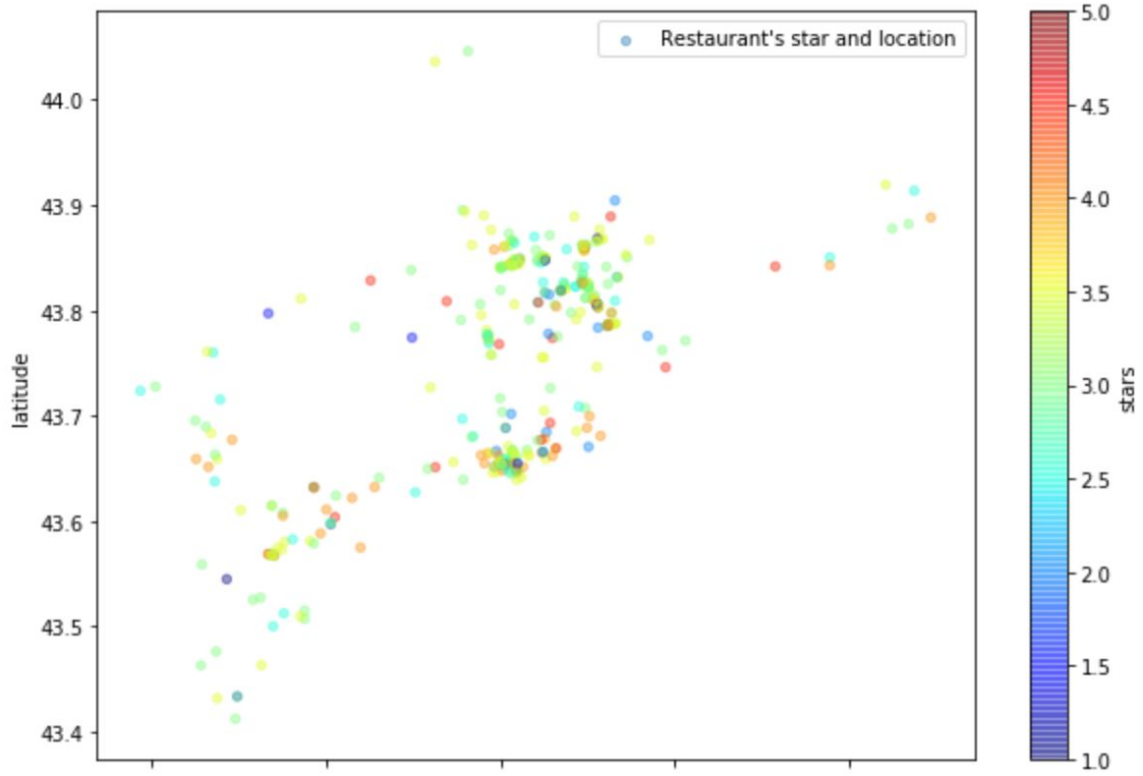
In [22]:
```
#The vast majority of Chinese restaurants near Toronto are concentrated
#The city of Toronto is well-understood as a population gathering place
#The relatively low rents and mature university districts have become
GTA_chinese['city'].value_counts().plot(kind='pie')
```

Out[22]: <matplotlib.axes._subplots.AxesSubplot at 0x126d10748>



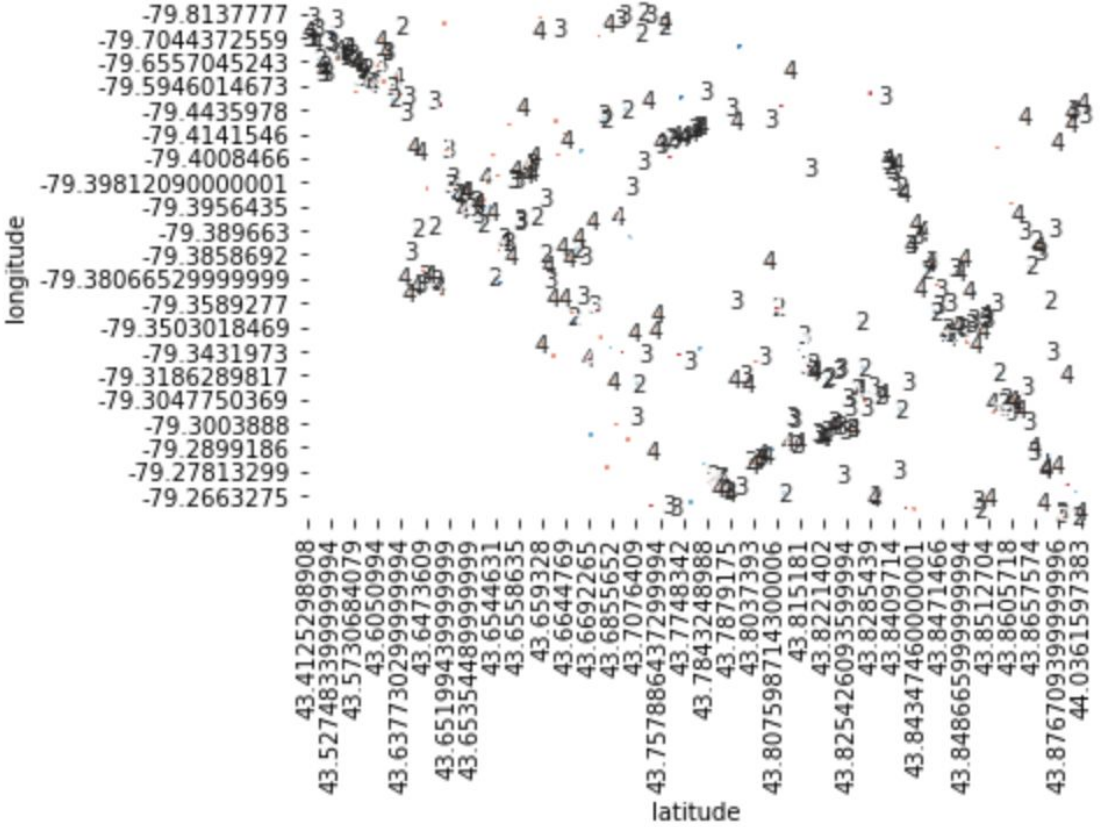
In [11]:
```
import matplotlib.pyplot as plt
import matplotlib.image as mpimg
#The graph show the distribution of the Chinese restaurant at greater toronto area with different star rat
#In the image, I used a latitude and longitude and a restaurant star to create a map of a Chinese restaura
GTA_chinese.plot(kind="scatter", x="longitude", y="latitude", alpha=0.4,colorbar=True,figsize=(10,7),c="st

plt.show()
```



In [37]:
```
sns.heatmap(GTA_chinese.groupby(['longitude', 'latitude'])['stars'].mean().unstack(),
            annot=True, cbar=False, fmt='.0f', cmap='RdBu_r')
```

Out[37]: <matplotlib.axes._subplots.AxesSubplot at 0x127ba8710>



Current Result

For Chinese restaurants at greater toronto area,

- Open rate = 83.6%> 74.9%
- Top area would be: Toronto downtown, Markham and Richmond Hill.
- Average star = 3.43 > 2.96
- Top streets: dundas street, yonge street
- Top Chinese restaurants: New Hong Kong, Luckee

Future steps

Step 1

Derived from the Toronto area to other regions, such as New York, Las Vegas and more. Compare the data horizontally and conclude the result.

Integrate data visualizations from different regions on the same map for a more intuitive understanding

Step 2

Further combine the content of review.csv and according to the id of the well Chinese restaurant, retrieve the evaluation of the restaurant by yelp users, and analyze the quality of the diners in the area.

Thank you